Signal Processing Research (SPR) Volume 4, 2015 www.seipub.org/spr
doi: 10.14355/spr.2015.04.004

Chinese Auto-Clustering of Oral Conversation Corpus Based on Contextual Features

Yue Chen*[1], Qi Chen[2], Minghu Jiang[3]

Lab of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084, China; College of Computer Science and Technology, Shandong University, Shandong, 250101, China; Lab of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084, China

*[1] yue-chen11@mails.tsinghua.edu.cn; [2] triplecq@gmail.com; [3] jiang.mh@tsinghua.edu.cn

Abstract

Chinese text clustering requires more linguistic knowledge in order to understand and analyze natural language accurately. To improve the accuracy of such clustering, in this article, we adopt SOM algorithm to add contextual features into the process of Chinese auto-clustering of oral corpus based on a contextual dictionary, and testify the effect of such a pragmatic application.

Keywords

Chinese Auto-Clustering; SOM; Oral Corpus; Contextual Features; Weight

Introduction

With more and more information filled in life, the thought of manual classification to explore data is impossible and are no doubt in need of the assistance of computers. One main branch of data mining is text clustering, which classifies texts by grouping objects. Chinese clustering requires word segmentation firstly, because Chinese words lack apparent boundaries like English. However, the accuracy of text clustering has never been high enough. Besides different algorithms, the most significant step is to involve deeper linguistic knowledge, and more and more linguistic features like syntax and semantics are adopted in the process, instead of merely relying on occurrence frequency.

However, the application of linguistics in clustering is restricted in some degree. From the perspective of NLP (Natural Language Processing), frequency may reveal some features, but is much less reliable when some highly occurring words are more prone to multi-meaning. Syntax analyses principles and process of sentences construction, and semantics focuses on the meaning, which both concentrate on the words and sentences themselves instead of considering the knowledge beyond the surface of texts.

Adding pragmatic features into clustering is inevitable in order to develop clustering. Pragmatics brings in and emphasizes speakers' part, which is out of the text itself[1]. Saussure once put forward two different conceptions: linguistics of language and linguistics of speaking, the latter of which indicates outside knowledge of the text as pragmatics. And context, though applied by other subjects like semantics as well, is a core part of pragmatics. One popular standard divides context into context of situation and context of culture. The former covers information of situation like time, place and subject; information of speakers and hearers such as identity, status, age, gender, job and their relationship; atmosphere and mood like privacy; motivation and effect and so on. While the later refers to common knowledge; specific knowledge of a specific target; and cultural background such as religion, habits and public convention[2]. While few predecessors ever touched this part, for the difficulty of the extraction and formalization of pragmatic features. This article trys to weight words with high contextual features based on contextual dictionary, in order to testify the effect of Chinese text clustering by reconstructing vector space model.

Related Works

Most text clustering methods concentrate on changing algorithms in order to accommodate the corpus, such as

hierarchical clustering and k-means. New types of corpuses are also involved into analysis, like WEB texts and newly derived short texts.

In the region of clustering involving linguistics, syntax and semantics are previously applied to clustering. In 2004, semantic analysis was almost firstly adopted in Chinese text clustering by establishing vector space model of term weight based on the theory of Latent Semantic Analysis (LSA)[3], following many other analysis concentrating on semantic and conceptual application to reduce dimensionality and weight effective parts at the same time. Other researchers analyse syntactic similarities of similar cluster, like Wang Man and Jiang[4]. While all these analysis concentrate on the surface of texts; only three scholars take a sip of pragmatics features apparently.

Wu Yun covers the recognition of semantic colours and inclinations, which though absorbed by semantics these years, are still important parts of pragmatics[5]. Wang Man and Jiang Minghu's research, which refers to the extraction and modelling of oral dialogue's features, apparently considers pragmatic feature as a new direction of clustering in the future[4]. The most relatively concentrated try was Chen Baojian, who regards contextual features as an independent element. He re-clusters texts by weighting pragmatic words based on disintegrating three pragmatic features from every chosen word and triple certain words' frequency[6]. However, his data only cover three themes (hospital, school and restaurant), which are even unreliable, and his method as well as criteria for extraction are still controversial and disputed without authority.

Experiment

Materials

We adopt oral dialogue corpus of overseas postgraduates majored in Chinese in Tsinghua University, most of whom are good at Chinese. The corpus is divided into 21 subjects (hospital, restaurant, ticket booking, inside school, after school, airport, book store, bank, market, tourism, visa service, beauty salon, post office, party, renting houses, photo studio, asking the way, interview, renting cars, insurance and transportation), with each consisting of 50 valued dialogues, in total of 1050 dialogues.

We also introduce a contextual dictionary (Fig.1): the Oxford-Duden Pictorial English-Chinese Dictionary, which 'paints' words in specific situations with different themes. Since the arrangement of the whole dictionary is based on context, we use this dictionary as a criterion of choosing contextual words.

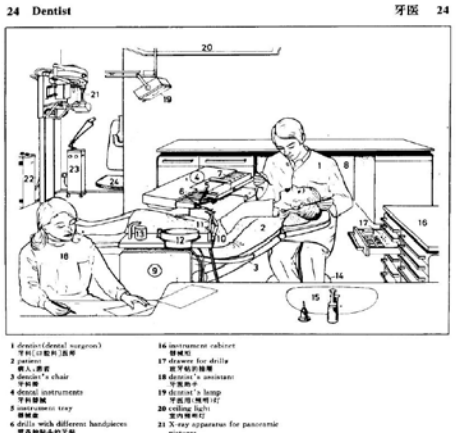

FIG. 1 SIMPLE OF CONTEXTUAL DICTIONARY

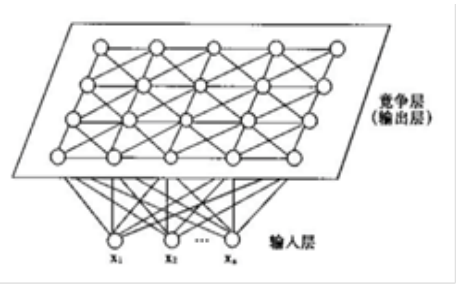

FIG. 2 SOM STRUCTURE

Algorithm and Method

We adopt SOM(self-organization map), firstly raised by T. Kohonen in 1981[7], which involves input layer to input corpus vectors and output layer to cluster different "nerve cells" as different subjects through an algorithm (see Fig.2). This method is unique because it resembles the biology nerve system.

The nerve cell j and Ij in (1) represents the input signal. While the Wij is the weight between nerve cell "i" to nerve cell "j". The output activity Yj is as below in the equation (2):

$$I_j = \sum_i W_{ij} X_i \tag{1}$$

$$\frac{dY_j}{dt} = I_j + \sum_{k \in S_i} r_k Y_j - g(Y_j) \tag{2}$$

Corpus Pre-Processing

Before processing, we correct errors, remove the qualifications, input the dictionary's theme-related words into Excel, conduct word segmentation using ICTCLAS and occurrence frequency of each dialogue and theme by ANTCONC after removing stop words. It turns out that only 8 themes can be adopted based on the dictionary, which are hospital(A), restaurant(B), renting houses(C) , inside school(D), after school(E), airport(F), beauty salon(G), and bank(H).

Procedure

The experiment is divided into two parts. The first part is called "default". We cluster those 8 subjects automatically, and record different training epochs such as 10, 200, 500, and 1500 to analyse their effects.

The second one is called "dictionary", we choose 10 contextual words from the dictionary which also occur mostly under each theme. For example, "酒精(medical alcohol), 血(blood), 药(medicine), 纱布(gauze), 医保卡(Medicare card), 手术(operation), 外科(surgical department), 内科(medical department), 病(sickness), 刀(operating knife)", occur both frequently under the theme of "hospital" in the dictionary as well as under the same theme in original corpus. As a result, there are 80 contextual words, and then we rerun the clustering after these words are weighted twice in vector model. Records are also saved under the same training epochs as group "default.

In this comparison, we regard "dictionary" as an improved vector based on chosen contextual words, which play as the modelling of contextual features.

Data Analysis

As the diagrams below, Fig.3 is the effect of the "default" group, when it runs 10 epochs; Fig.4 is the effect of the "dictionary" group when it trains 200 epochs. Each cell on the right side represents the number of texts clustered together under certain subject, and the distances between each cell is illustrated in the left diagrams. The perfect result should be 50 each cell and the distances between subjects are as far as possible, which means each subject cluster accurate texts from all conversations.

Firstly, the result of dictionary group is better than the default one. In default's group, numbers are 32, 35, 39, 42, 49, 59, 63 and 81, while the latter one's are 39, 44, 47, 49, 51, 56 and 65. Not only the numbers between 40 and 60 are more (5) in the latter than the default (3), but also the range of dictionary group (26) is much narrower than the former (49), nearly a half. Numbers in each cell are closer to 50 in Figure.4, which means those conversations are clustered much more evenly.

To show this effect apparently, we further increase the nerve cells into 100 as in Fig.5. The left diagram represents "default" group, while the right stands for "dictionary". We can clearly recognize the blank cells in "dictionary" are more than those in "default", and texts on the right are more concentrated. It reveals the better effect of contextual features visually.

Secondly, the distances between points in Fig.5 are much broader than those in Fig.4. It reveals that clustering is improved in terms of the distances between different subjects are further and different clusters are more outstanding as well as recognizable.

Lastly, it is obvious that there is a cell of a much higher number than 50, like the cell of 81 in Fig.4. This may result from that dialogues in "inside school"(D) and "after school"(E) are of great similarity, because there are similar words such as "teacher", "students", and "classes", which results in a convergence. This ambiguous phenomenon refers to a further problem of the definition and extraction of distinct contextual features, and the classification of different subjects.

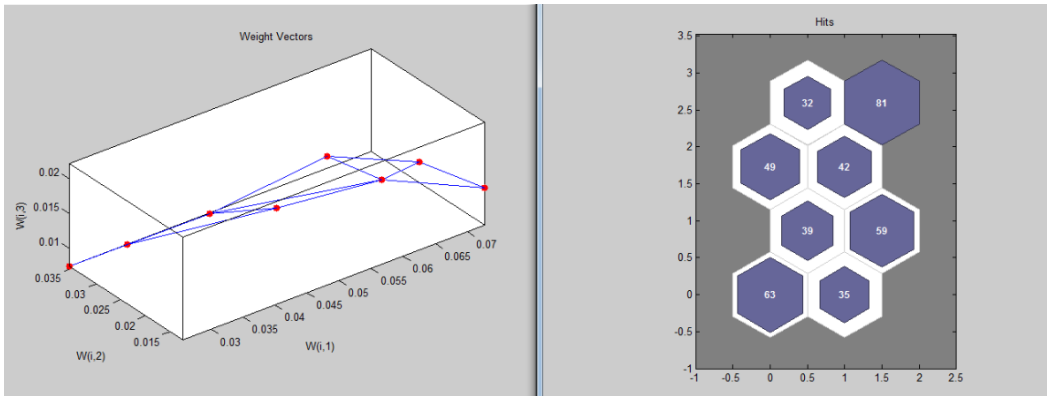

FIG. 3 "DEFAULT", E10

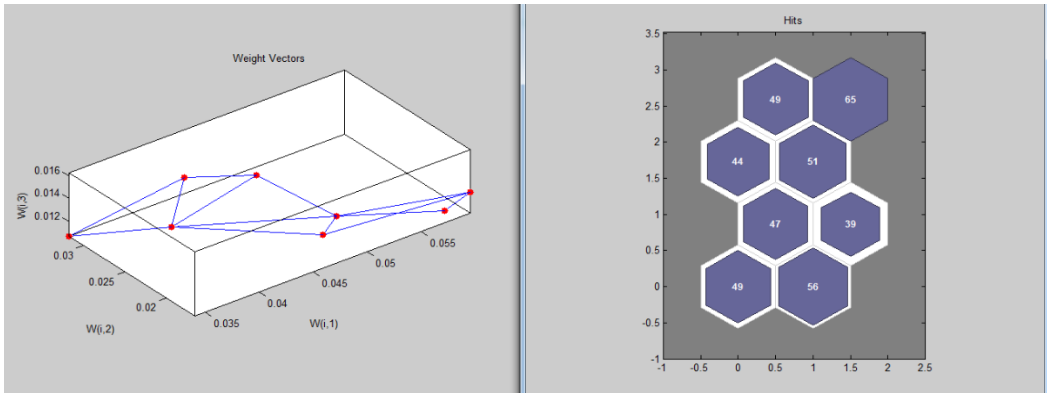

FIG. 4 "DICTIONARY", E10

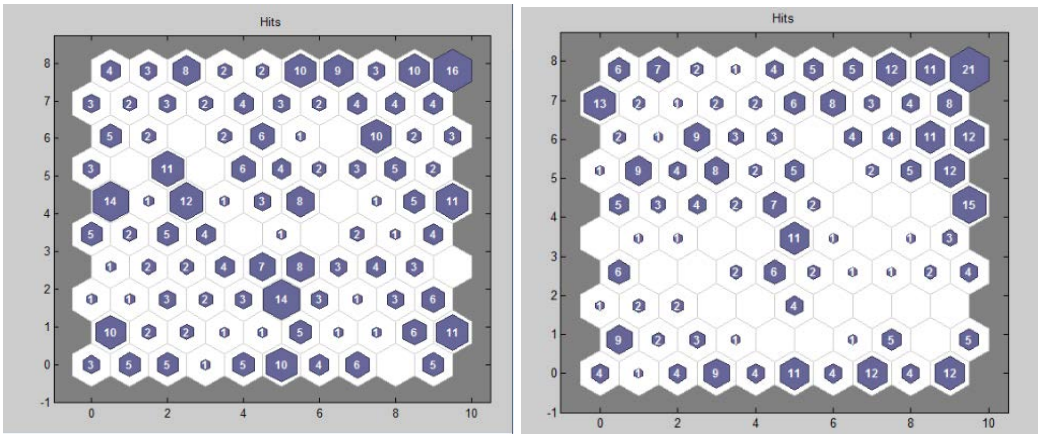

FIG. 5 "DEFAULT" AND "DICTIONARY" IN 100 CELLS, E10

To sum up, we can come to a conclusion based on sufficient evidences that contextual features do have a positive effect on Chinese clustering, and the dictionary is reliable as a criterion for the extraction of pragmatic, especially contextual features.

Discussion and Future

Although contextual features do have positive effects on clustering according to this experiment, the effect is not that obvious and distinct. One reason might be the mistakes in choosing words. Because we use a manual method to check apparently related words from the dictionary, other words which might be crucial to clustering might be ignored by us. Additionally, the dictionary itself has some problems as well, though it bases on contextual background, it concerns formal expressions mostly and covers so many professional words like "control buttons for the pacemaker unit", which makes the contextual words mostly nouns. Such limitations may affect the true effect of contextual features.

Furthermore, since the dictionary only covers 8 related subjects, the rest 13 subjects are left wasted. Further

analysis should be taken so that all 21 themes could be clustered and testified. Although few authoritative criteria could be used to extract contextual features, manual work based on pragmatic knowledge might be reliable. So our next step is to choose contextual words by detaching contexture features of highly frequent words under each theme according to detailed and numerous pragmatic concerns (like 10-20 criteria, so that words with highly contexture features could be recognized). If such attempt can achieve good results, it can not only improve the accuracy of Chinese clustering, but also analyse contextual features of words with new criteria.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Fund (61171114) and Key Fund (61433015), and National Social Science Major Fund (14ZDB154 & 13ZD187) of China.

REFERENCES

[1] Yule, G., Pragmatics. 1996: Oxford University Press

[2] Zhenyu Suo, Pragmatic Course book, 2000, Peking University Press.

[3] Guojun Ma, Weiguo Yun, Researching of Chinese Text Clustering Based on Latent Semantic Index, Modern Electronic Technique, 2005. 28(10): No.58-59.

[4] Man Wang, Minghu Jiang, Extraction and Modeling of Oral Corpus Features based on Cerebral Organization Mechanism, Collection in memory of 80th birthday of William S.-Y.Wang, No.19.

[5] Yun Wu, Words' Semantic Orientation Analysis Based on Text Orientation Indentification, 2008, Beijing University of Post and Telecommunications (Beijing), No.64.

[6] Baojian Chen, Auto-Clustering of Conversation Corpus Based on Contextual Features, 2014, Tsinghua University (Beijing), No.73

[7] Kohonen, T., Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 1982. (1)(43): p. 59–69